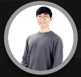
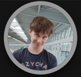
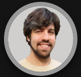
NAVER
NAVER LABS
Europe
University of Sussex

Compositional preference models for aligning LMs


Dongyoung Go
dongyoung.go@navercorp.com
@dongyoung4091

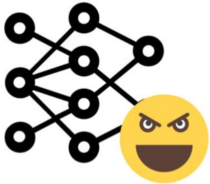

Tomek Korbak
http://tomekkorbak.com
@tomekkorbak


Germán Kruszewski
german.kruszewski@naverlabs.com
@germank


Jos Rozen
jos.rozen@naverlabs.com
@josrzn


Marc Dymetman
marc.dymetman@gmail.com
@MarcDymetman

Problem: Aligning language models with human preferences

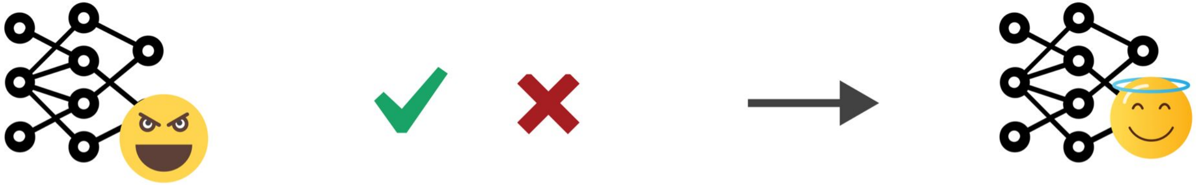


pretrained LM a(x)

✓ ✗

preference data

Problem: Aligning language models with human preferences



pretrained LM a(x)

preference data

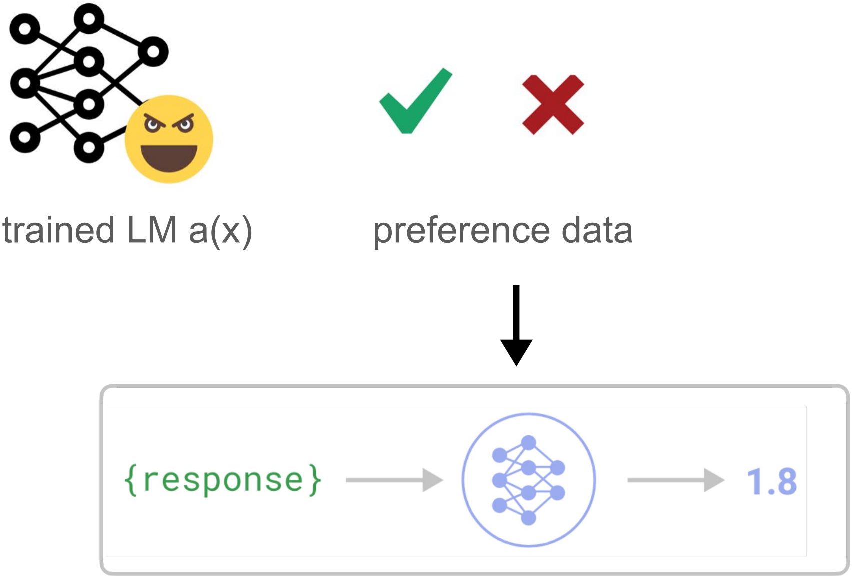
aligned LM π(x)

The goal is to fine-tune a pretrained LM $a(x)$,
so that the fine-tuned LM $\pi(x)$ incorporates some preferences

Dominant approach: Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback

Step1. Training a preference model (PM) to predict human preference judgments



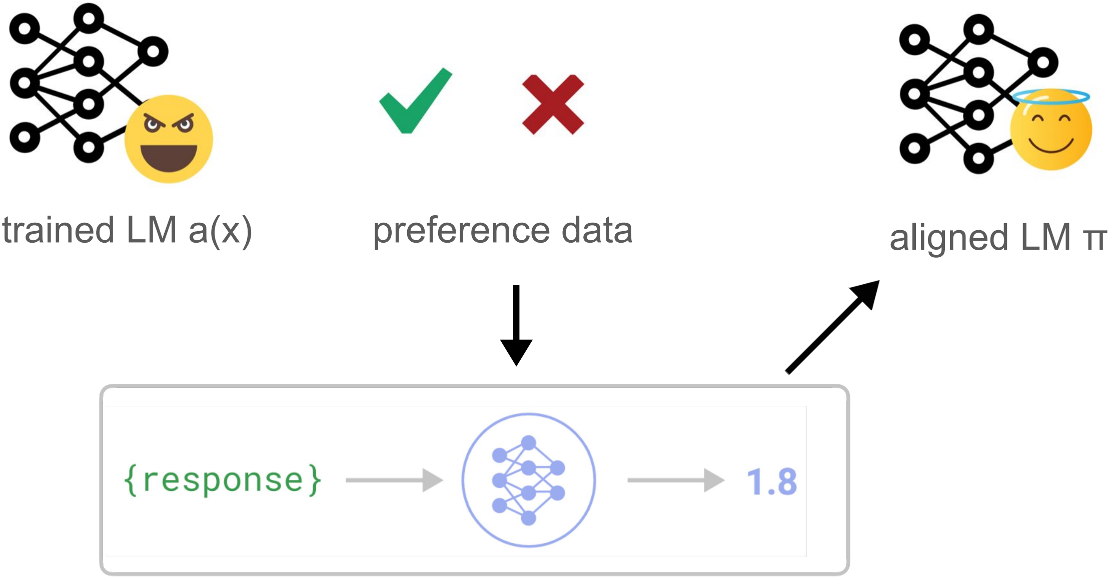
pretrained LM a(x)

preference data

preference model

Reinforcement Learning from Human Feedback

Step2. Finetuning an LM to maximize the reward given by the PM.



pretrained LM a(x)

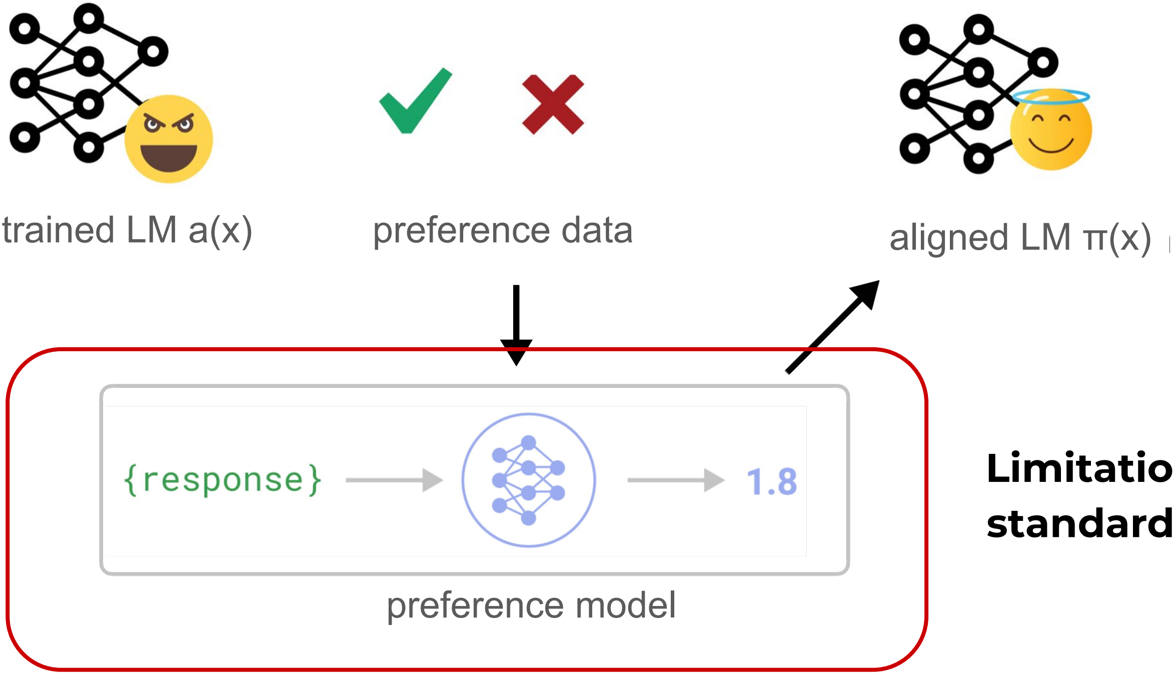
preference data

aligned LM π(x)

preference model

Reinforcement Learning from Human Feedback

Step2. Finetuning an LM to maximize the reward given by the PM.



pretrained LM a(x)

preference data

aligned LM π(x)

{response}

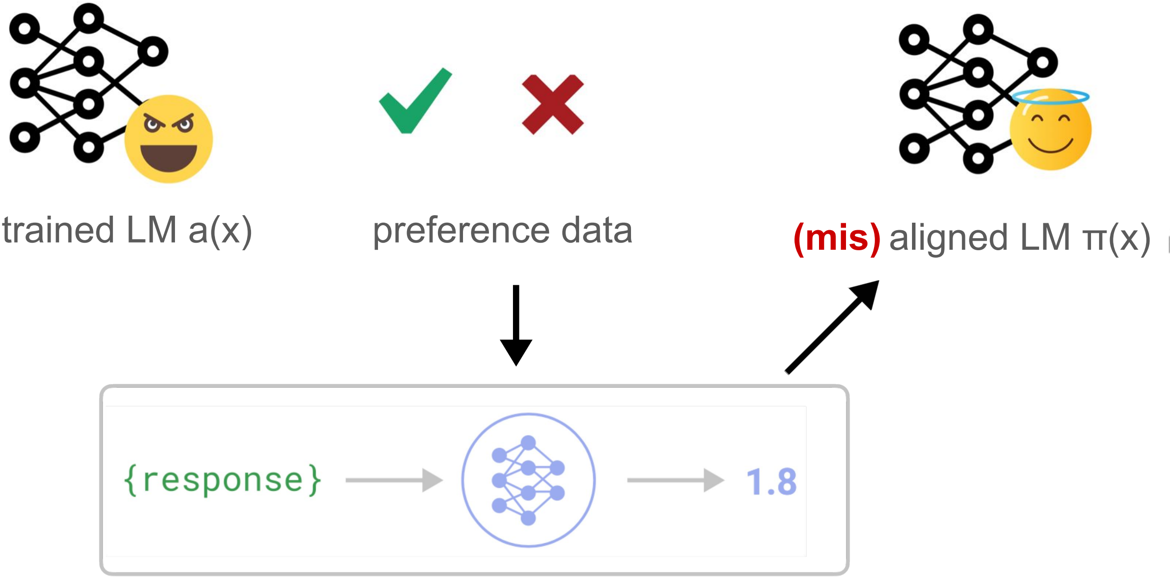
1.8

preference model

Limitation of standard PM

Limitation of standard PM

1. Susceptible to overfitting the preference dataset (Overoptimization)



pretrained LM a(x)

preference data

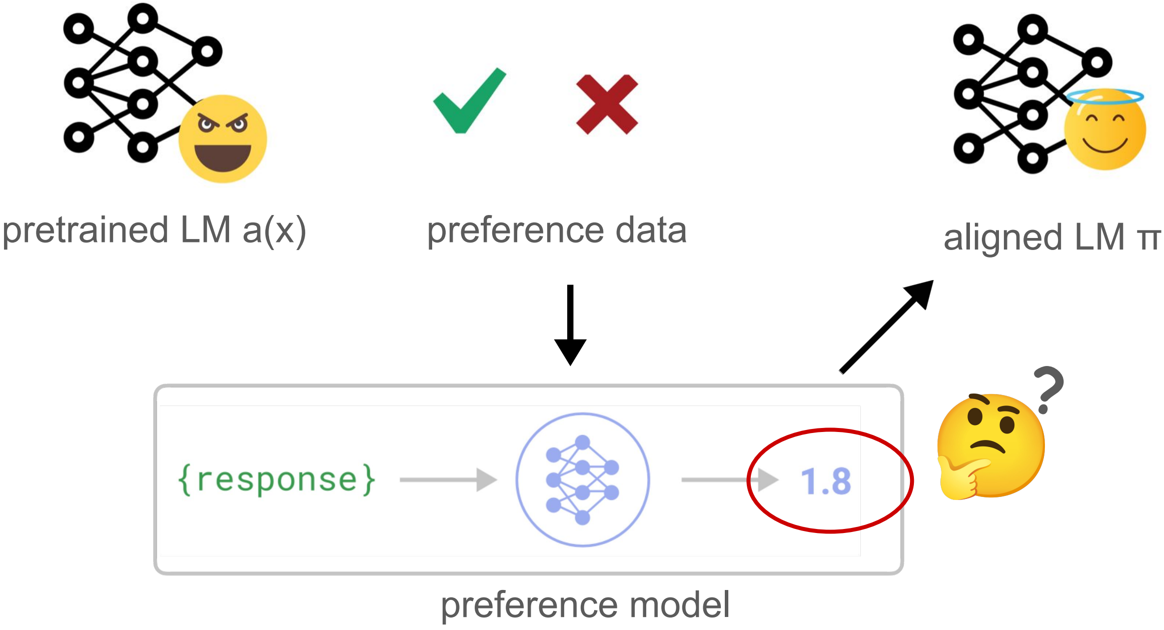
(mis) aligned LM π(x)

{response} → 1.8

(Overfitted) preference model

Limitation of standard PM

2. Difficult to interpret and to oversee



pretrained LM a(x)

preference data

aligned LM π(x)

preference model

Compositional Preference Model (CPM)

Simple yet effective framework for learning PM that is

1. More robust to overoptimization
2. More transparent and interpretable
3. More aligned with desired preference

by providing inductive bias from **human insight**
combine with **LM capabilities**

How compositional preference models work?

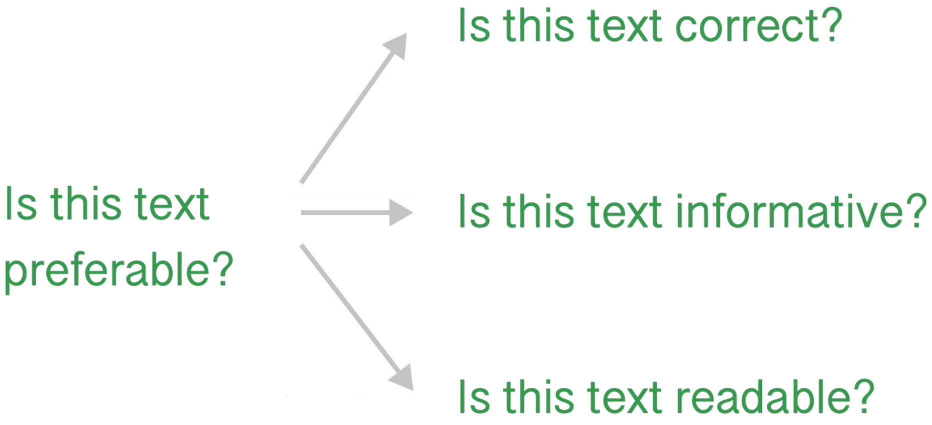
Step1. Feature Decomposition

Step2. Feature Scoring

Step3. Aggregation

How compositional preference models work?

Step 1: Feature decomposition



Is this text preferable?

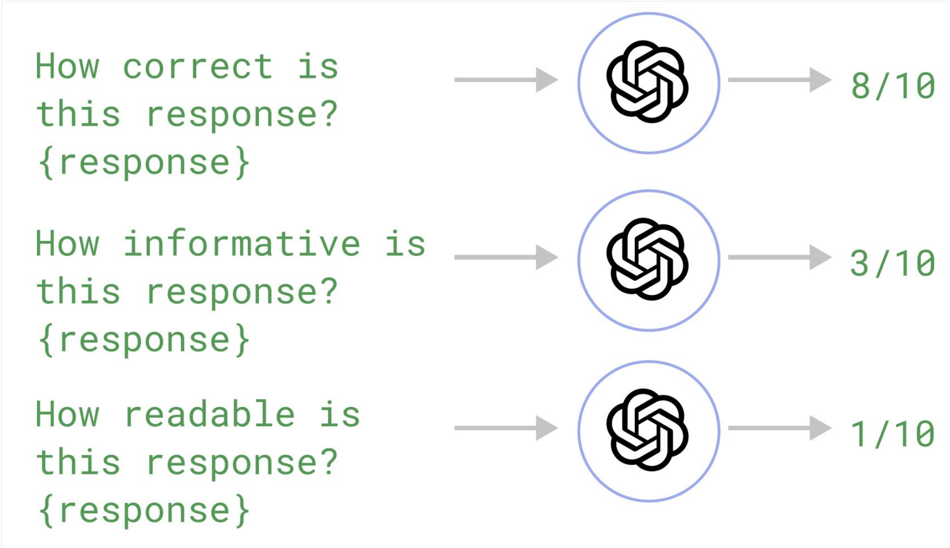
Is this text correct?

Is this text informative?

Is this text readable?

How compositional preference models work?

Step 2: Feature scoring



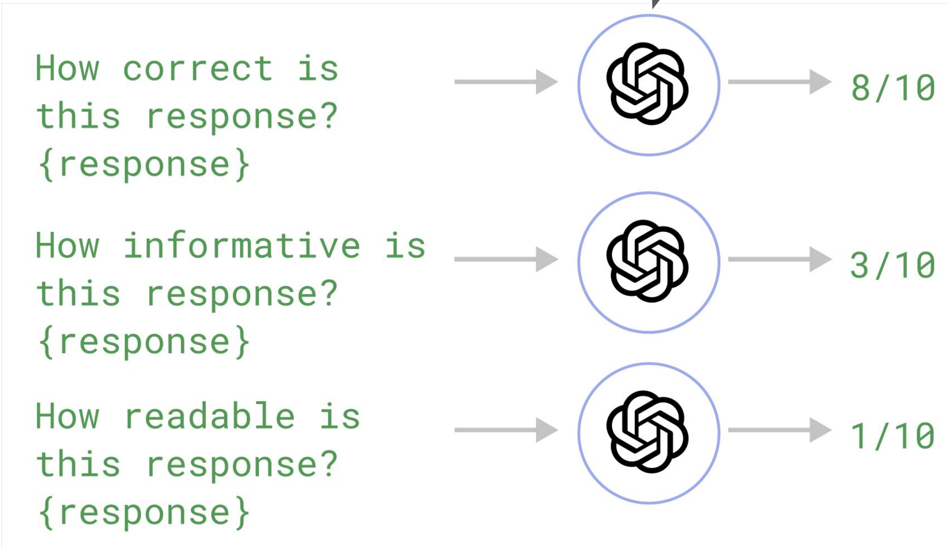
How correct is this response? {response} → 8/10

How informative is this response? {response} → 3/10

How readable is this response? {response} → 1/10

How compositional preference models work?

Step 2: Feature scoring



How correct is this response? {response}

8/10

How informative is this response? {response}

3/10

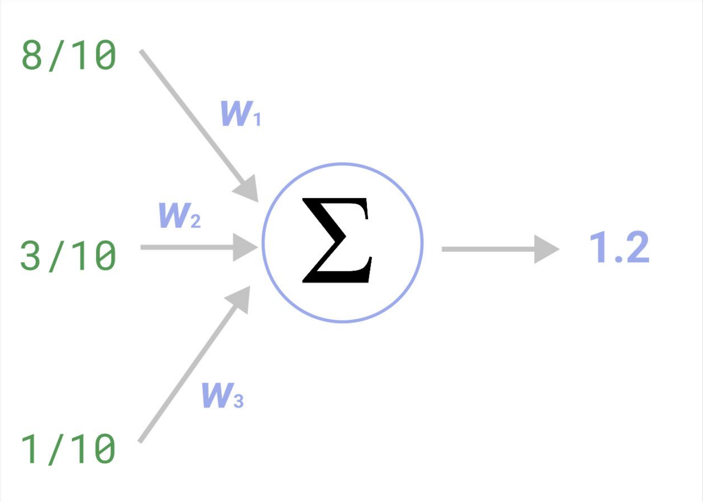
How readable is this response? {response}

1/10

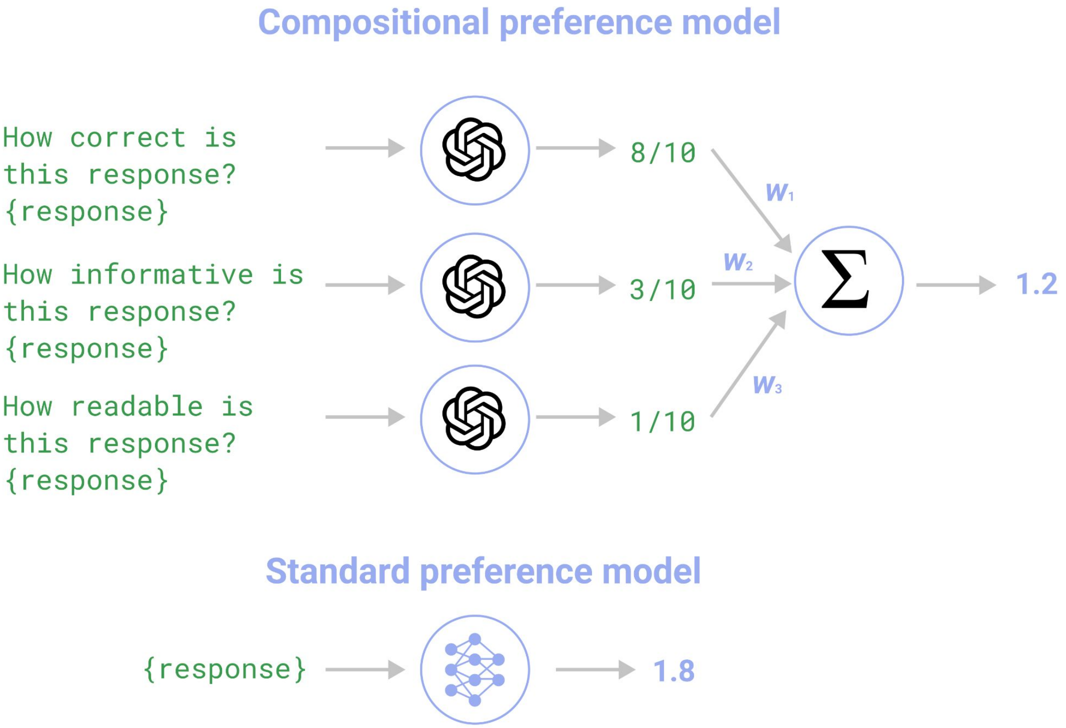
Enable simpler model for the following steps!

How compositional preference models work?

Step 3: Aggregation



Why compositional preference models works this way?



Compositional preference model

How correct is this response? {response}

How informative is this response? {response}

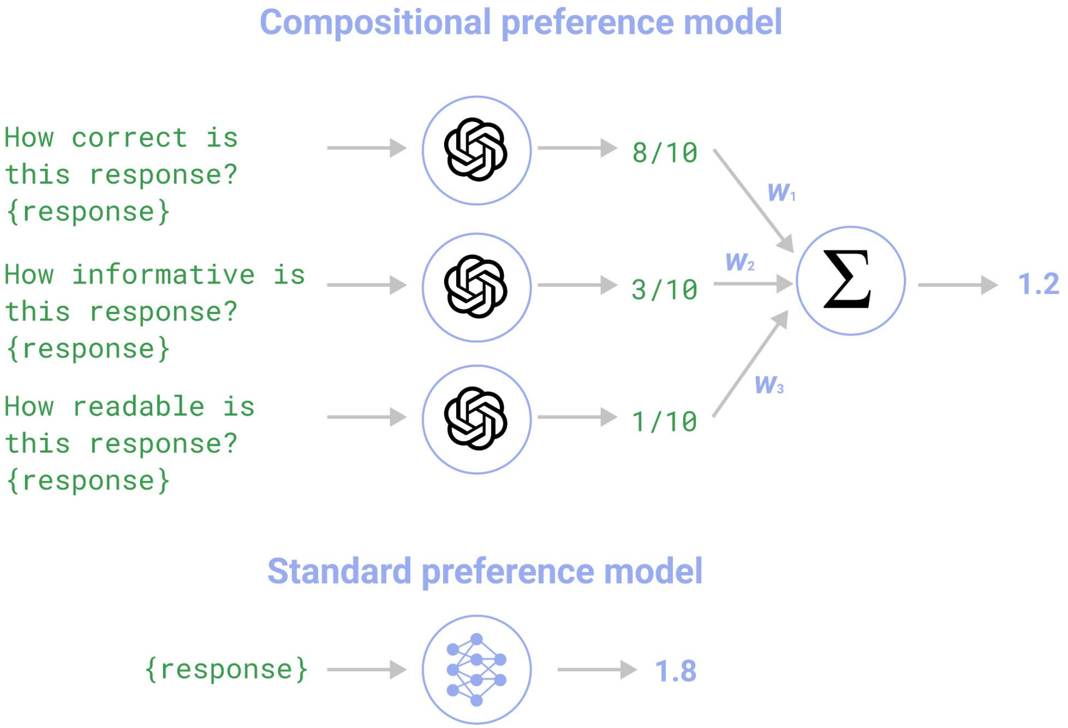
How readable is this response? {response}

Standard preference model

{response}

CPMs are given the human prior knowledge about which features determine preferences

Why compositional preference models works this way?



Compositional preference model

`How correct is this response? {response}` → 8/10 → w_1

`How informative is this response? {response}` → 3/10 → w_2

`How readable is this response? {response}` → 1/10 → w_3

Σ → 1.2

Standard preference model

`{response}` → 1.8

CPMs are given the human prior knowledge about which features determine preferences

This provides interpretable inductive bias and limits their susceptibility to overfitting

Experiment setting

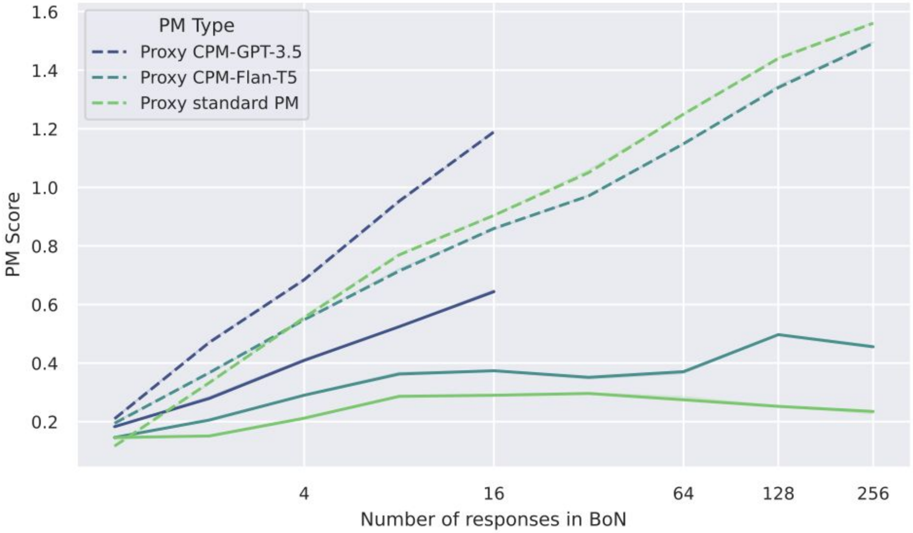
Dataset: HH-RLHFdataset, SHP dataset

Features for CPM: 13 features (helpfulness, specificity, intent, factuality, easy-to-understand, relevance, readability, enough-detail, biased, fail-to-consider-individual-preferences, repetitive, fail-to-consider-context and too-long)

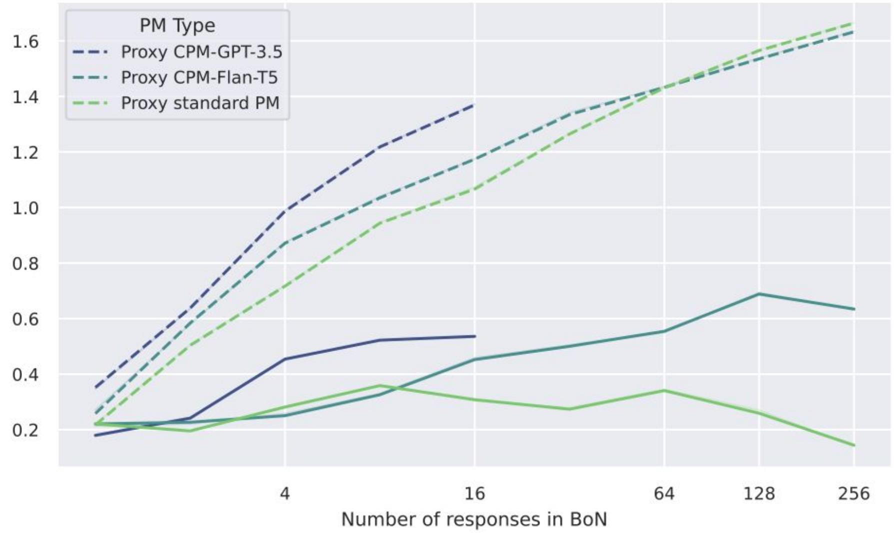
Model: Flan-T5-XL (3B parameters) for both of conventional PM and CPM feature extractor (GPT-3.5 is also explored for ablation)

Experiment - Robustness to Overoptimization

HH-RLHF dataset

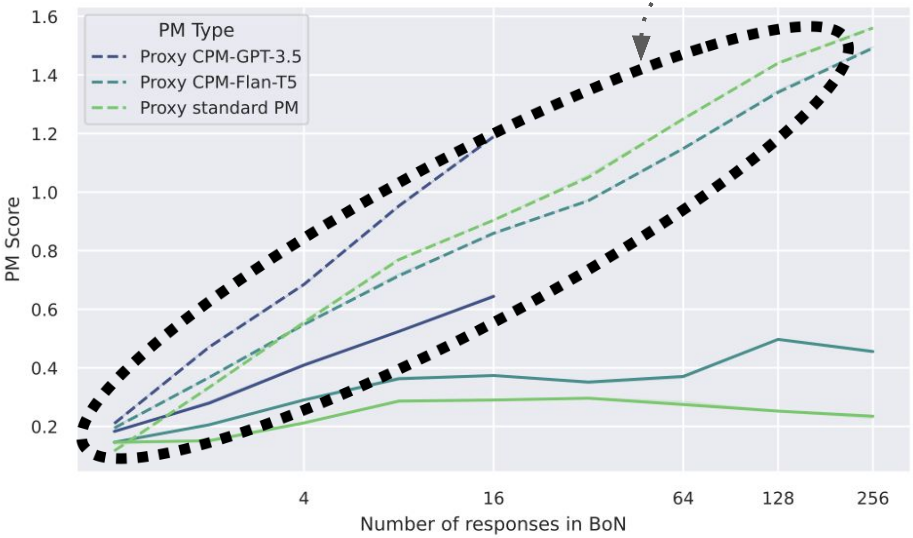


SHP dataset



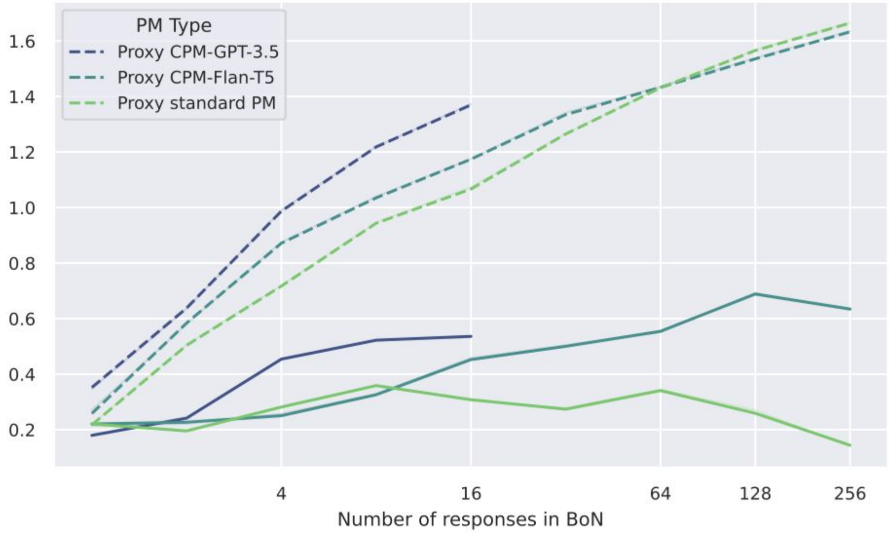
Experiment - Robustness to Overoptimization

HH-RLHF dataset



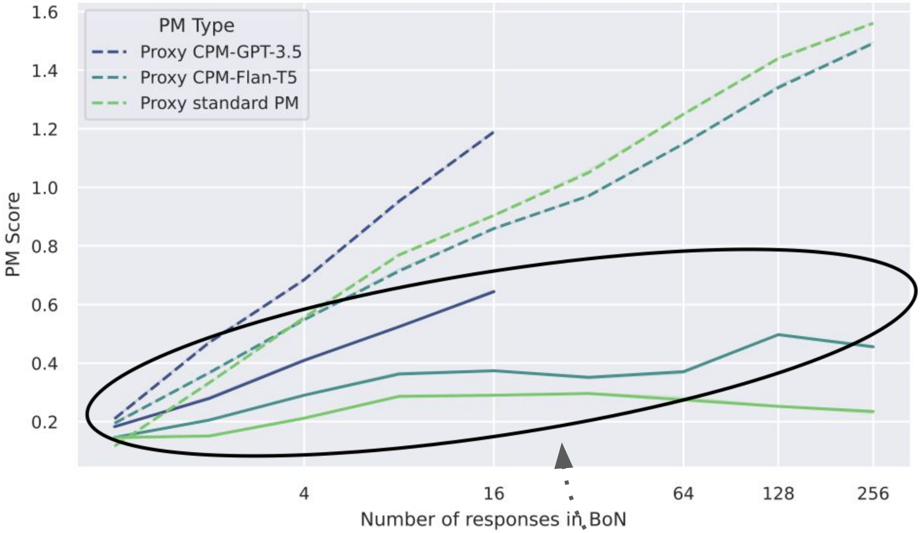
Proxy Model

SHP dataset



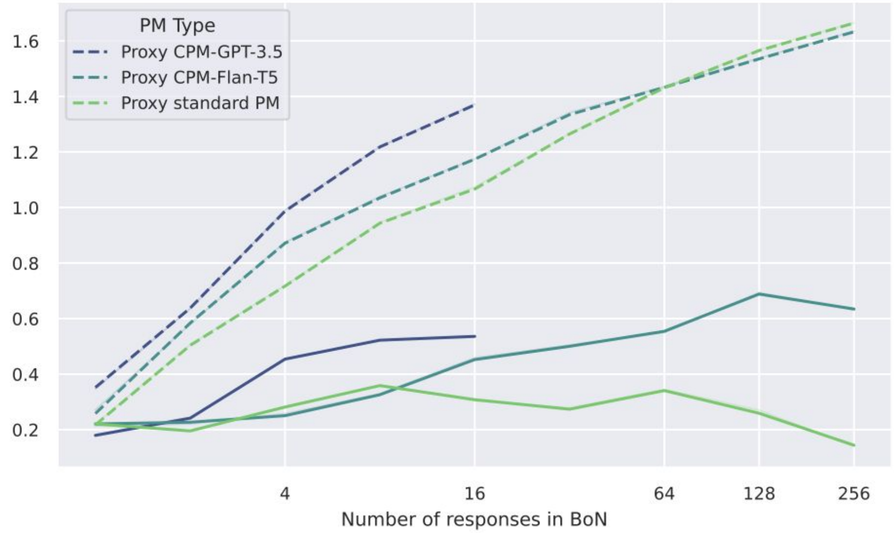
Experiment - Robustness to Overoptimization

HH-RLHF dataset



Gold Model

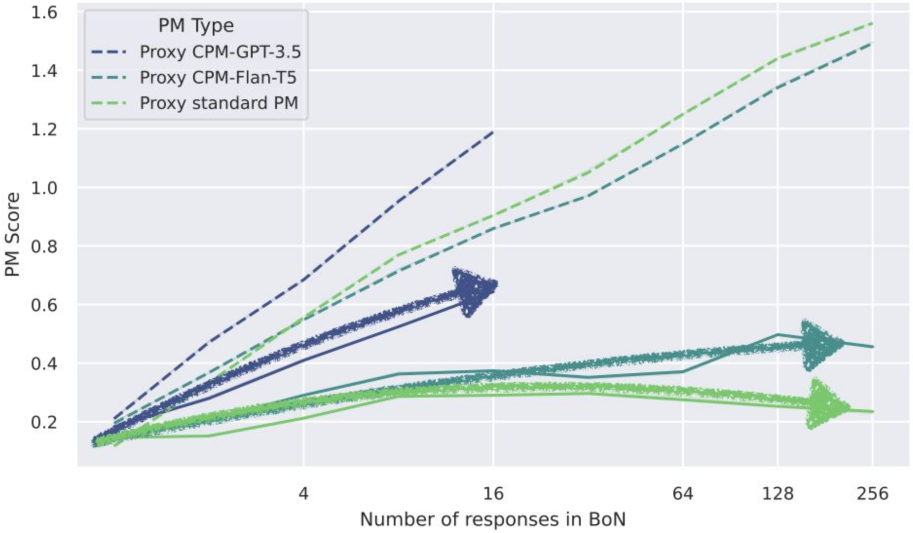
SHP dataset



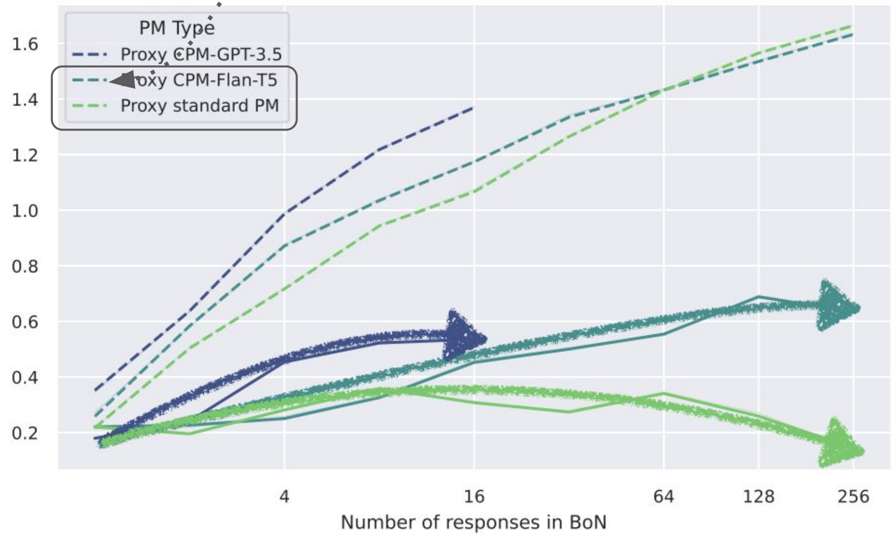
Experiment - Robustness to Overoptimization

Same Model!
(Flan-T5)

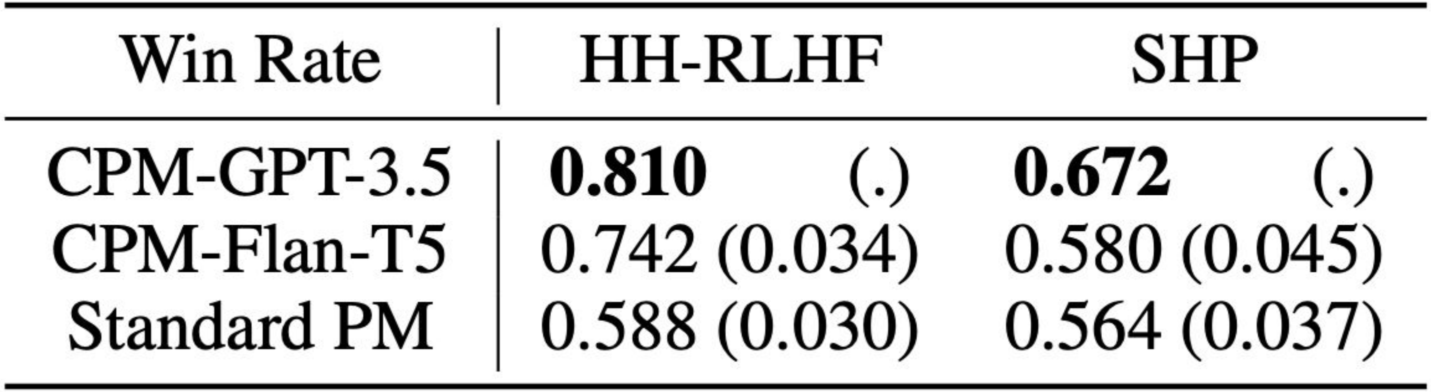
HH-RLHF dataset



SHP dataset



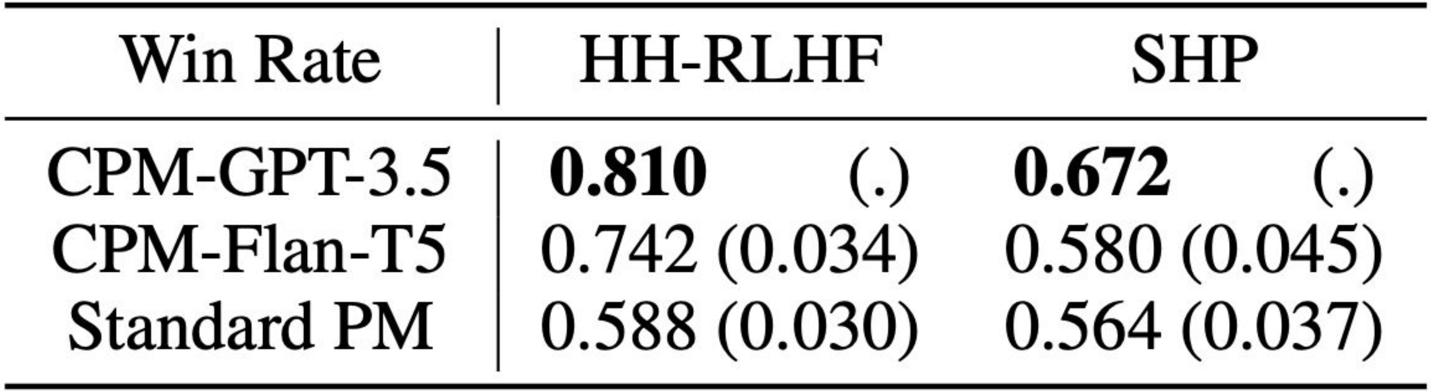
Experiment - Alignment (LM evaluator, Claude-2)



Win Rate	HH-RLHF	SHP
CPM-GPT-3.5	**0.810** (.)	**0.672** (.)
CPM-Flan-T5	0.742 (0.034)	0.580 (0.045)
Standard PM	0.588 (0.030)	0.564 (0.037)

Table: Win rate over initial generation after BoN sampling based on each PM. Except CPM-GPT-3.5, we independently conduct 10 rounds of BoN(n = 16) samplings and report the average win rate along with standard error.

Experiment - Alignment (LM evaluator, Claude-2)

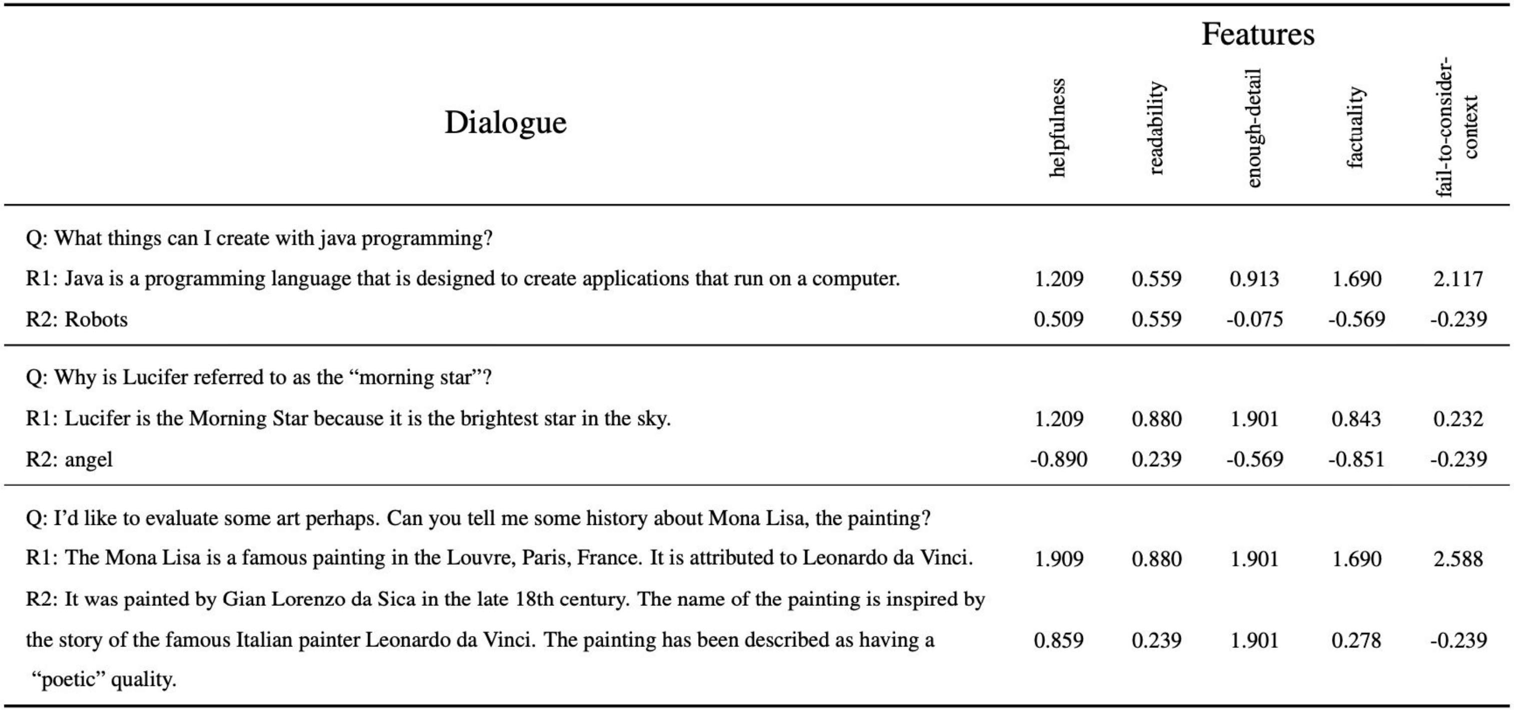


Win Rate	HH-RLHF	SHP
CPM-GPT-3.5	**0.810** (.)	**0.672** (.)
CPM-Flan-T5	0.742 (0.034)	0.580 (0.045)
Standard PM	0.588 (0.030)	0.564 (0.037)

Table: Win rate over initial generation after BoN sampling based on each PM. Except CPM-GPT-3.5, we independently conduct 10 rounds of BoN(n = 16) samplings and report the average win rate along with standard error.

prior knowledge injected into a CPM is robustly helpful for learning human preferences!

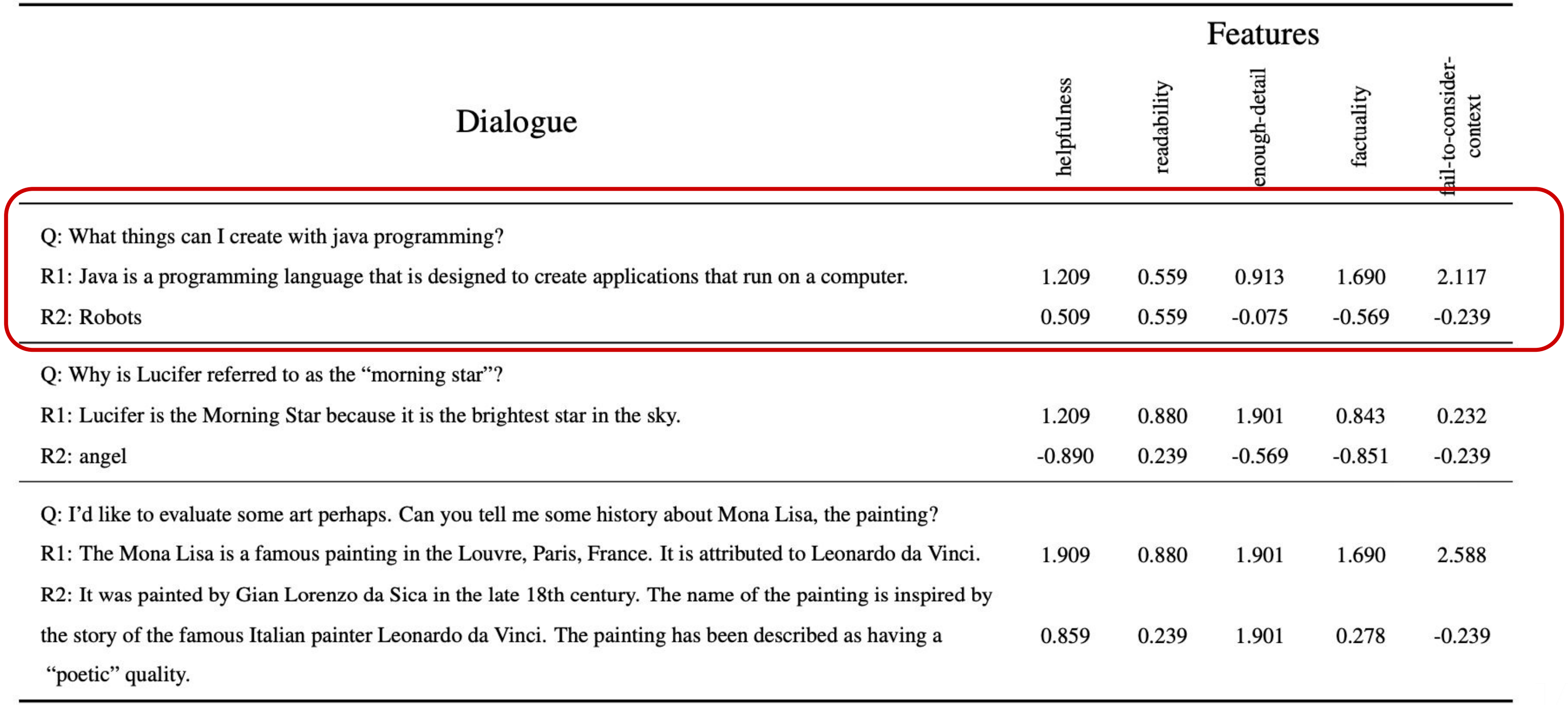
Experiment - Model Interpretability



Dialogue	Features				
	helpfulness	readability	enough-detail	factuality	fail-to-consider-context
Q: What things can I create with java programming?					
R1: Java is a programming language that is designed to create applications that run on a computer.	1.209	0.559	0.913	1.690	2.117
R2: Robots	0.509	0.559	-0.075	-0.569	-0.239
Q: Why is Lucifer referred to as the “morning star”?					
R1: Lucifer is the Morning Star because it is the brightest star in the sky.	1.209	0.880	1.901	0.843	0.232
R2: angel	-0.890	0.239	-0.569	-0.851	-0.239
Q: I’d like to evaluate some art perhaps. Can you tell me some history about Mona Lisa, the painting?					
R1: The Mona Lisa is a famous painting in the Louvre, Paris, France. It is attributed to Leonardo da Vinci.	1.909	0.880	1.901	1.690	2.588
R2: It was painted by Gian Lorenzo da Sica in the late 18th century. The name of the painting is inspired by the story of the famous Italian painter Leonardo da Vinci. The painting has been described as having a “poetic” quality.	0.859	0.239	1.901	0.278	-0.239

Pre-selected features are easily interpretable by its definition!

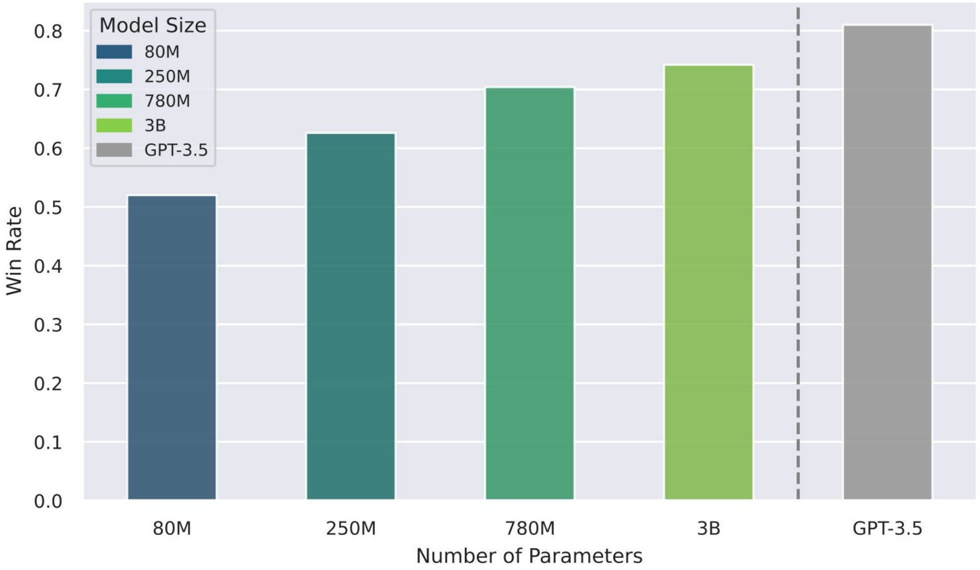
Experiment - Model Interpretability



Dialogue	Features				
	helpfulness	readability	enough-detail	factuality	fail-to-consider-context
Q: What things can I create with java programming?					
R1: Java is a programming language that is designed to create applications that run on a computer.	1.209	0.559	0.913	1.690	2.117
R2: Robots	0.509	0.559	-0.075	-0.569	-0.239
Q: Why is Lucifer referred to as the "morning star"?					
R1: Lucifer is the Morning Star because it is the brightest star in the sky.	1.209	0.880	1.901	0.843	0.232
R2: angel	-0.890	0.239	-0.569	-0.851	-0.239
Q: I'd like to evaluate some art perhaps. Can you tell me some history about Mona Lisa, the painting?					
R1: The Mona Lisa is a famous painting in the Louvre, Paris, France. It is attributed to Leonardo da Vinci.	1.909	0.880	1.901	1.690	2.588
R2: It was painted by Gian Lorenzo da Sica in the late 18th century. The name of the painting is inspired by the story of the famous Italian painter Leonardo da Vinci. The painting has been described as having a "poetic" quality.	0.859	0.239	1.901	0.278	-0.239

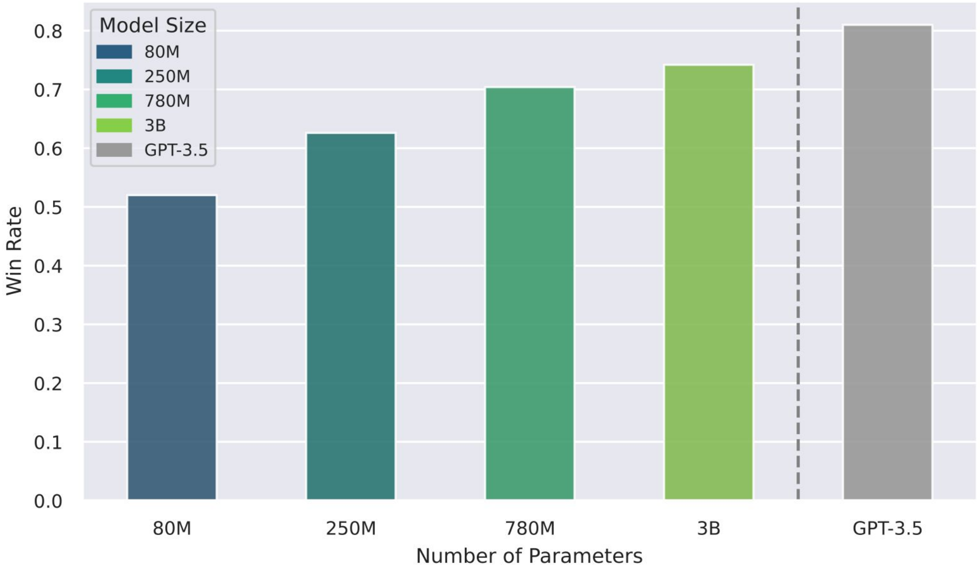
Pre-selected features are trivially interpretable by its definition!

Experiment - Scaling law



Gradually increase this size from Flan-T5 “small” (80M) to “XL” (3B)

Experiment - Scaling law



Gradually increase this size from Flan-T5 “small” (80M) to “XL” (3B)

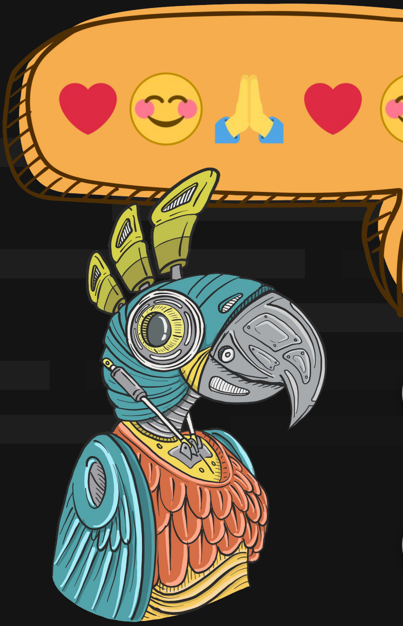
Win rate steadily improve with increasing LM size

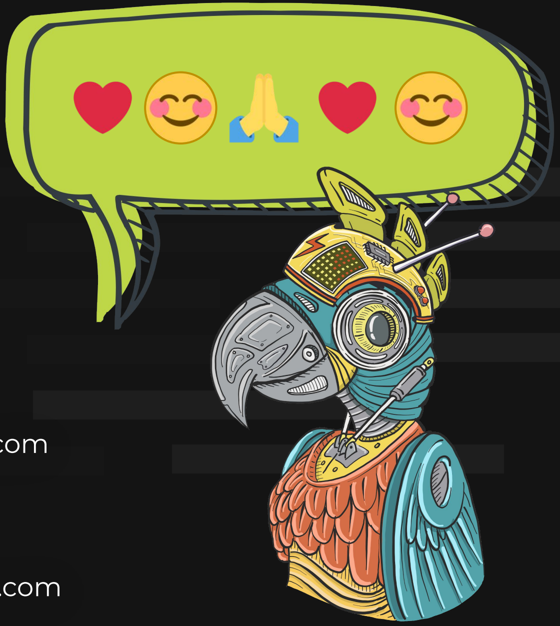
CPMs can become even more useful as extractor LMs become more capable

Conclusion

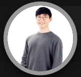
- Intricacy of feature extraction can be delegated to LLM
- Human prior can be used to guide the feature dimension
- CPM is interpretable, robust and overseeable PM
- Potential for the scalable oversight of models with superhuman capabilities.

NAVER NAVER LABS Europe University of Sussex

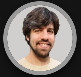




Thank you!

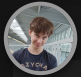


Dongyoung Go
dongyoung.go@navercorp.com
@dongyoung4091



Germán Kruszewski
https://germank.github.io
@germank

Marc Dymetman
marc.dymetman@gmail.com
@MarcDymetman



Tomek Korbak
http://tomekkorbak.com
@tomekkorbak

Jos Rozen
jos.rozen@naverlabs.com
@josrzn